

# SINGLE-CROSSOVER DYNAMICS: FINITE VERSUS INFINITE POPULATIONS

ELLEN BAAKE AND INKE HILDEBRANDT

**ABSTRACT.** Populations evolving under the joint influence of recombination and resampling (traditionally known as genetic drift) are investigated. First, we summarise and adapt a deterministic approach, as valid for infinite populations, which assumes continuous time and single crossover events. The corresponding nonlinear system of differential equations permits a closed solution, both in terms of the type frequencies and via linkage disequilibria of all orders. To include stochastic effects, we then consider the corresponding finite-population model, the Moran model with single crossovers, and examine it both analytically and by means of simulations. Particular emphasis is on the connection with the deterministic solution. If there is only recombination and every pair of recombined offspring replaces their pair of parents (i.e., there is no resampling), then the *expected* type frequencies in the finite population, of arbitrary size, equal the type frequencies in the infinite population. If resampling is included, the stochastic process converges, in the infinite-population limit, to the deterministic dynamics, which turns out to be a good approximation already for populations of moderate size.

## 1. INTRODUCTION

It is well known that the dynamics of populations under recombination is notoriously difficult to treat, even if the population is so large that stochastic fluctuations due to genetic drift can be neglected, so that the time evolution may be described by a deterministic dynamical system. The difficulty lies in the inherent nonlinearity of the corresponding (difference or differential) equations, which stems from the interaction of pairs of parental individuals during recombination.

The overwhelming part of the literature on the dynamics of recombination deals with *discrete* time. The first solutions go back to Geiringer (1944, [15]) and Bennett (1954, [8]); their construction was more recently worked out in greater detail and completeness by Dawson [11]. The underlying mathematical structures were investigated within the framework of genetic algebras, see [18, 19], or [23]. Quite generally, the solution relies on a certain nonlinear transformation (known as Haldane linearisation) from (gamete or haplotype) frequencies to suitable linkage disequilibria, which decouple from each other and decay geometrically. But if more than three loci are involved, this transformation must be constructed via recursions that involve the parameters of the recombination process, and is not available explicitly except in the trivial case of free recombination (i.e., independent gene loci). A different approach was taken by Barton and Turelli [7] by translating type frequencies into certain sets of moments; the resulting iterations are well-suited for symbolic manipulation, numerical treatment, and biological interpretation, but do not lead to (and do not primarily aim at) explicit solutions. For a review of the area, see [9, Ch. V.4].

It has turned out recently [3, 6, 5] that, in contrast to the discrete-time situation, the corresponding dynamics in *continuous* time permits a simple explicit solution in a biologically relevant special case, namely, the situation in which at most one crossover happens at any

given time. Here, only recombination events may occur that partition the sites of a sequence (or the loci on a chromosome) into two *ordered* subsets that correspond to the sites before and after a given crossover point. In contrast to previous approaches, the solution is given directly in terms of the original type frequencies; but, again, a transformation to certain linkage disequilibria is available that linearise and diagonalise the system. The main simplification lies in the fact that the transformation is independent of the recombination rates and is available in a simple explicit form.

However, the focus in modern population genetics is on *finite* populations, which also experience genetic drift (that is, resampling). In order to investigate the relevance of the deterministic solution for the more involved stochastic system, we shall, in this paper, examine the corresponding finite population model, namely, the Moran model with recombination, and compare some of its properties with the infinite population model, both analytically and by means of simulations. We shall show the following:

- In a *finite* population with recombination, but without resampling (i.e., every pair of recombined offspring replaces its pair of parents), the type frequencies are, *in expectation*, given by the corresponding quantities for *infinite* populations. This property holds exactly as long as the genetic material is conserved in the recombination process, and to a very good approximation otherwise. This connection between stochastic and deterministic models is by no means obvious: It is usually reserved for populations of individuals that evolve independently (like branching processes); or to systems with interactions that do not change the expected type frequencies (such as the Moran model with independent mutation and resampling). In contrast, interaction due to recombination *does* change the expected composition of the population, and does, therefore, not appear to be of this simple form.
- If the joint action of recombination and resampling is considered, the process of type frequencies differs from the deterministic dynamics, even in expectation, particularly if the population is small. It follows from general arguments, however, that the former converges to the latter when population size goes to infinity. By means of simulations, we show that this limit yields a good approximation of realistic biological situations: Convergence is fast enough to justify the use of the deterministic solution already for populations of moderate size (of the order of  $10^5$  individuals, say).

The paper is organised as follows. We first recapitulate the single-crossover differential equation and its solution. This is appropriate since the original articles [5, 6] use a rather abstract measure-theoretical framework suitable for very general type spaces, which is not easily accessible to theoretical biologists. We reduce the formalism to the important special case of finite type spaces, which is free from technical subtleties and permits an explicit representation in terms of probability vectors as well as illustration by means of concrete examples. We then change gears and explore the Moran model with recombination. With the help of standard techniques for Markov chains in continuous time, we develop the results stated above.

## 2. THE DETERMINISTIC APPROACH

**2.1. The model and its recombinator formulation.** We start by extracting from [5] and [6] the continuous-time dynamics of single crossovers in an infinite population. Let us consider genes or chromosomes as linear arrangements of sites, indexed by the set  $S := \{0, 1, \dots, n\}$ ; sites may be interpreted as either nucleotide positions in a stretch of DNA, or gene loci on a

chromosome. For each site  $i \in S$ , there is a set  $X_i$  of “letters” (to be interpreted as nucleotides or alleles, respectively) that may possibly occur at that site. To allow for a convenient notation, we restrict ourselves to the simple but important case of *finite* sets  $X_i$ ; for the full generality of arbitrary locally compact  $X_i$ , the reader is referred to the original articles.

A *type* is thus defined as a sequence  $x = (x_0, x_1, \dots, x_n) \in X_0 \times X_1 \times \dots \times X_n =: X$ , where  $X$  shall be called the *type space*. By construction,  $x_i$  is the  $i$ -th coordinate of  $x$ , and we define  $x_I := (x_i)_{i \in I}$  as the collection of coordinates with indices in  $I$ , where  $I$  is a subset of  $S$ . Types may be understood as alleles (if sites are nucleotide positions) or haplotypes (if sites are gene loci). We shall think at the haploid level, speak of gametes as individuals, and describe a population as a distribution of (absolute) frequencies (or a non-negative measure)  $\omega$  on  $X$ . Namely,  $\omega(\{x\})$  denotes the frequency of type  $x \in X$ , and  $\omega(A) := \sum_{x \in A} \omega(\{x\})$  for  $A \subset X$ ; we abbreviate  $\omega(\{x\})$  as  $\omega(x)$ . The set of all frequency distributions on  $X$  is denoted by  $\mathcal{M}$ . If we define  $\delta_x$  as the point measure on  $x$  (i.e.,  $\delta_x(y) = \delta_{x,y}$  for  $x, y \in X$ ), we can also write  $\omega = \sum_{x \in X} \omega(x) \delta_x$ . We may, alternatively, interpret  $\delta_x$  as the basis vector of  $\mathbb{R}_{\geq 0}^{|X|}$  that corresponds to  $x$  (where a suitable ordering of types is implied, and  $|X|$  is the number of elements in  $X$ );  $\omega$  is thus identified with a vector in  $\mathbb{R}_{\geq 0}^{|X|}$ .

At this stage, frequencies need not be normalised;  $\omega(x)$  may simply be thought of as the size of the subpopulation of type  $x$ , measured in units so large that it may be considered a continuous quantity. The corresponding normalised version  $p := \omega / \|\omega\|$  (where  $\|\omega\| := \sum_{x \in X} \omega(x) = \omega(X)$  is the total population size) is then a probability distribution on  $X$ , and may be identified with a probability vector.

Recombination acts on the links between the sites; the links are collected into the set  $L := \{\frac{1}{2}, \frac{3}{2}, \dots, \frac{2n-1}{2}\}$ . We shall use Latin indices for the sites and Greek indices for the links, and the implicit rule will always be that  $\alpha = \frac{2i+1}{2}$  is the link between sites  $i$  and  $i+1$ ; see Figure 1.

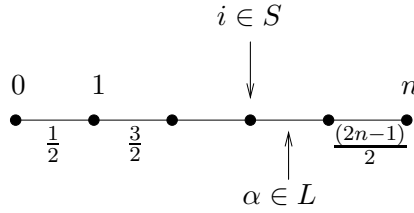


FIGURE 1. Sites and links.

We shall make the simplifying assumption that only single crossovers occur at a time (which is realistic if linkage is tight, i.e., if the sites belong to the DNA sequence of, say, less than a megabase, or represent a few adjacent loci). More precisely, for every  $\alpha \in L$ , every individual exchanges, at rate  $\varrho_\alpha/2$ , the sites after link  $\alpha$  with those of a randomly chosen partner. Explicitly, if the ‘active’ and the partner individual are of type  $x$  and  $y$ , then the new pair has types  $(x_0, x_1, \dots, x_{[\alpha]}, y_{[\alpha]}, \dots, y_n)$  and  $(y_0, y_1, \dots, y_{[\alpha]}, x_{[\alpha]}, \dots, x_n)$ , where  $[\alpha]$  ( $\lceil \alpha \rceil$ ) is the largest integer below  $\alpha$  (the smallest above  $\alpha$ ); see Fig. 2. Since every individual can occur as either the ‘active’ individual or as its randomly chosen partner, we have a total rate of  $\varrho_\alpha$  for crossovers at link  $\alpha$ . For later use, we also define  $\varrho := \sum_{\alpha \in L} \varrho_\alpha$ . The dynamics of the type

frequencies is then given by the system of differential equations

$$(1) \quad \dot{\omega}_t(x) = \sum_{\alpha \in L} \varrho_\alpha \left( \frac{1}{\|\omega_t\|} \omega_t(x_0, \dots, x_{[\alpha]}, *, \dots, *) \omega_t(*, \dots, *, x_{[\alpha]}, \dots, x_n) - \omega_t(x) \right)$$

for all  $x \in X$ , where a “\*” at site  $i$  stands for  $X_i$  and denotes marginalisation over the letters at site  $i$  (i.e.,  $\omega_t(x_0, \dots, x_{[\alpha]}, *, \dots, *)$  is the frequency of individuals which have letters  $x_0, \dots, x_{[\alpha]}$  at the sites before  $\alpha$ , and arbitrary letters from  $X_{[\alpha]}, \dots, X_n$  at the sites after  $\alpha$ ).

Note that the assumption of a reciprocal exchange (between two parents that combine into *two* offspring under conservation of the genetic material) is an ‘effective’ description that results from taking the symmetry of recombination into account. A more detailed model starts out from two parents combining into *one* offspring; but the symmetry of the process again leads to Eq. 1 and is thus equivalent to our picture, see [9, Ch. II.2.1]. In contrast to the deterministic situation, however, these details do play a role in finite populations; this will be taken up in Sec. 3.

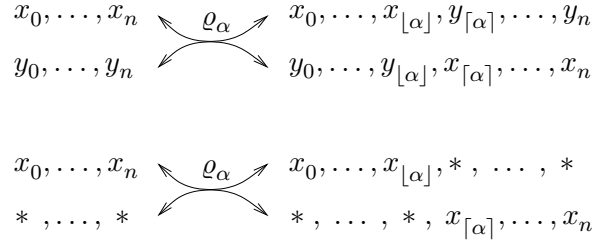


FIGURE 2. Upper panel: Recombination between individuals of type  $x$  and  $y$ . Lower panel: The corresponding ‘marginalised’ version that summarises all events by which individuals of type  $x$  are gained or lost. Note that, in either case, the process can go both ways, as indicated by the arrows.

We have retained the non-normalised version for the frequencies here for easier comparison with the stochastic model (the stochastic process will come naturally as a family of integer-valued random variables, which sum up to the population size  $N$ ). Of course, the more familiar normalised version emerges if  $\omega_t$  in (1) is replaced by  $p_t := \omega_t / \|\omega_t\|$ ; the normalising factor on the right-hand side is then, of course, unity.

Note that the model (1) implies Hardy-Weinberg proportions (i.e., frequencies of diplotypes are given by independent combination of the corresponding haplotypes at all times). In continuous time, this only applies exactly if the duration of the diplophase is negligible. However, it is a good approximation if recombination is rare at the scale of an individual’s lifetime.

An important ingredient to the solution of the large, nonlinear system of differential equations (1) lies in its reformulation in terms of recombination operators. Let us introduce the projection operators  $\pi_i$ ,  $i \in S$ , via

$$(2) \quad \begin{aligned} \pi_i : X_0 \times X_1 \times \dots \times X_n &\rightarrow X_i \\ (x_0, x_1, \dots, x_n) &\mapsto x_i, \end{aligned}$$

i.e.,  $\pi_i$  is the canonical projection to the  $i$ -th coordinate. Likewise, for any index set  $I \subset S$ , one defines a projector  $\pi_I : X \rightarrow X_I := \bigtimes_{i \in I} X_i$  via  $(x_0, x_1, \dots, x_n) \mapsto (x_i)_{i \in I} =: x_I$ . We shall frequently use the abbreviations  $\pi_{<\alpha} := \pi_{\{1, \dots, [\alpha]\}}$  and  $\pi_{>\alpha} := \pi_{\{[\alpha], \dots, n\}}$ , as well as  $x_{<\alpha} := \pi_{<\alpha}(x)$ ,  $x_{>\alpha} := \pi_{>\alpha}(x)$ . The projectors  $\pi_{<\alpha}$  and  $\pi_{>\alpha}$  may be thought of as *cut and*

*forget* operators because they take the leading or trailing segment of a sequence  $x$ , and forget about the rest.

Whereas the  $\pi_I$  act on the types, we also need the induced mapping at the level of the population, namely,

$$(3) \quad \begin{aligned} \pi_I : \mathcal{M} &\longrightarrow \mathcal{M} \\ \omega &\longmapsto \pi_I.\omega := \omega \circ \pi_I^{-1}, \end{aligned}$$

where  $\pi_I^{-1}$  denotes the preimage under  $\pi_I$ . The operation  $.$  (where the dot is on the line and should not be confused with a multiplication sign) is known as the “pullback” of  $\pi_I$  w.r.t.  $\omega$ ; in terms of coordinates, the definition may be spelled out as

$$(\pi_I.\omega)(x_I) = \omega \circ \pi_I^{-1}(x_I) = \omega(\{x \in X \mid \pi_I(x) = x_I\}), \quad x_I \in X_I.$$

So,  $\pi_I.\omega$  is nothing but the marginal distribution of  $\omega$  with respect to the sites in  $I$ . In particular,  $(\pi_{<\alpha}.\omega)(x_{<\alpha}) = \omega(x_0, x_1, \dots, x_{[\alpha]}, *, \dots, *)$  is the marginal frequency of sequences prescribed at the sites before  $\alpha$ , and vice versa for the sites after  $\alpha$ . Note that, although we have defined  $\pi_I.$  as acting on nonnegative measures (or vectors) only, this linear operator may be extended to the set of all measures.

Now, recombination (at the level of the population) means the relinking of a randomly chosen leading segment with a randomly chosen trailing segment. We therefore introduce (elementary) recombination operators (or *recombinators*, for short),  $R_\alpha : \mathcal{M} \rightarrow \mathcal{M}$  for  $\alpha \in L$ , defined by

$$(4) \quad R_\alpha(\omega) := \frac{1}{\|\omega\|} ((\pi_{<\alpha}.\omega) \otimes (\pi_{>\alpha}.\omega)).$$

Here, the tensor product reflects the independent combination (i.e., the product measure) of the two marginals  $\pi_{<\alpha}.\omega$  and  $\pi_{>\alpha}.\omega$ .  $R_\alpha$  is therefore a *cut and relink* operator.  $R_\alpha(\omega)$  may be understood as the population that emerges if *all* individuals of the population  $\omega$  disintegrate into their leading and trailing segments, and these are relinked randomly; the sites before  $\alpha$  are then in linkage equilibrium with respect to those after  $\alpha$ . Note that  $\|R_\alpha(\omega)\| = \|\omega\|$ .

In terms of coordinates, Eq. (4) reads

$$(5) \quad (R_\alpha(\omega))(x) = \frac{1}{\|\omega\|} \omega(x_0, x_1, \dots, x_{[\alpha]}, *, \dots, *) \omega(*, \dots, *, x_{[\alpha]}, \dots, x_n).$$

The recombination dynamics (1) may thus be compactly rewritten as

$$(6) \quad \dot{\omega}_t = \sum_{\alpha \in L} \varrho_\alpha (R_\alpha(\omega_t) - \omega_t) = \sum_{\alpha \in L} \varrho_\alpha (R_\alpha - \mathbb{1})(\omega_t) =: \Phi(\omega_t),$$

where  $\mathbb{1}$  is the identity operator.

**2.2. Solution via recombinators.** The solution of (6) relies on some elementary properties of our recombinators. Most importantly, they are idempotents and commute with each other, i.e.,

$$(7) \quad R_\alpha^2 = R_\alpha, \quad \alpha \in L,$$

$$(8) \quad R_\alpha R_\beta = R_\beta R_\alpha, \quad \alpha, \beta \in L.$$

These properties are intuitively very plausible: if linkage equilibrium has already been established at link  $\alpha$ , then further recombination at that link does not change the situation; and if a product measure is formed with respect to two links  $\alpha$  and  $\beta$ , the result does not depend

on the order in which the links are affected. For the proof, we refer to [6, Prop. 2]; let us only mention here that it relies on the elementary fact that, for  $\omega \in \mathcal{M}$ ,

$$\begin{aligned} \pi_{<\alpha} \cdot (R_\beta(\omega)) &= \pi_{<\alpha} \cdot \omega && \text{for } \beta \geq \alpha, \text{ and} \\ \pi_{>\alpha} \cdot (R_\beta(\omega)) &= \pi_{>\alpha} \cdot \omega && \text{for } \beta \leq \alpha; \end{aligned}$$

that is, recombination at or after  $\alpha$  does not affect the marginal frequencies at sites before  $\alpha$ , and vice versa.

We now define *composite* recombinators as

$$R_G := \prod_{\alpha \in G} R_\alpha \quad \text{for } G \subset L.$$

Here, the product is to be read as composition; it is, indeed, a matrix product if the recombinators are written in their matrix representation, which is available in the case of finite types considered here (see [3]). By property (8), the order in the composition plays no role. Furthermore, (7) and (8) obviously entail  $R_G R_H = R_{G \cup H}$  for  $G, H \subset L$ .

The solution of the single-crossover dynamics (6) may now be given in closed form as

$$(9) \quad \omega_t = \sum_{G \subset L} a_G(t) R_G(\omega_0) =: \varphi_t(\omega_0)$$

with coefficients

$$a_G(t) = \prod_{\alpha \in L \setminus G} e^{-\varrho_\alpha t} \prod_{\beta \in G} (1 - e^{-\varrho_\beta t}),$$

and initial value  $\omega_0$ ; i.e.,  $\varphi_t$  is the *semigroup* belonging to the recombination equation (6). For the proof, the reader is referred to [6, Thm. 2], or [5, Thm. 3] (the former article contains the original, the latter a shorter and more elegant version of the proof). Let us only note here that the coefficients  $a_G(t)$  have the following intuitive explanation. In a given individual, consider the set of links that has, until  $t$ , been hit by at least one crossover event. The probability that this set is  $G$  equals the probability that all links in  $G$  have already been hit, while those in the complement of  $G$  have not; by independence across links, this probability is just  $a_G(t)$ .

We would like to note, however, that this plausible argument should not be taken too far: For example, an analogous solution, although suggestive, does *not* apply to the corresponding single-crossover model in *discrete* time (its solution is, in fact, much more difficult, see the discussion in [6]). Indeed, explicit solutions to large, nonlinear systems are rare – explicit semigroups are usually available for linear systems at best. For this reason, the recombination equation and its solution have already been taken up in the framework of functional analysis, where they have led to an extension of potential theory [22]. In the next paragraph, we will meet some further clues to an underlying linear structure that lurks behind the solution.

**2.3. Linkage disequilibria.** The approach described in the previous Section is somewhat unconventional from a population genetics perspective, in that it solves the recombination dynamics at the level of the type frequencies. As mentioned in the Introduction, however, one usually works with transformed quantities, like moments, cumulants, or linkage disequilibria (LDEs). Indeed, certain linkage disequilibria are intimately connected with the solution just presented, and provide the key to an underlying linearity, and hence simplicity, as we shall now see. Following [6], let us define what we shall call *LDE operators* via

$$(10) \quad T_G := \sum_{H \supset G} (-1)^{|H \setminus G|} R_H, \quad G \subset L,$$

where the underdot indicates the summation variable. Eq. (10) leads to the inverse  $R_G = \sum_{H \supset G} T_H$  by Möbius inversion, a versatile tool from combinatorial theory, see [1, Thm. 4.18]. It was shown in [6] that, if  $\omega_t$  is the solution (9), the transformed quantities  $T_G(\omega_t)$  satisfy

$$(11) \quad \frac{d}{dt} T_G(\omega_t) = - \left( \sum_{\alpha \in L \setminus G} \varrho_\alpha \right) T_G(\omega_t), \quad G \subset L,$$

which is a system of *decoupled, linear, homogeneous* differential equations, with solution  $T_G(\omega_t) = \exp(-\sum_{\alpha \in L \setminus G} \varrho_\alpha) T_G(\omega_0)$ . Note that this simple form emerged through the *nonlinear* transform (10) as applied to the solution of the *coupled, nonlinear* differential equation (6).

Obviously, the  $T_G(\omega_t)$  provide candidates for the definition of linkage disequilibria which decouple, and decay exponentially (unless, of course,  $G = L$ , which corresponds to the stationary state

$$(12) \quad \omega_\infty = T_L(\omega_0) = \frac{1}{\|\omega\|^{n-1}} \bigotimes_{i=1}^n (\pi_i \cdot \omega_0),$$

the population in linkage equilibrium with respect to all links). All that remains to be done is to identify a set of suitable components to work with (using all components of  $T_G(\omega_t)$ , for all  $G \subset L$ , would imply a lot of redundancy).

To this end, let us introduce the following shorthand notation. Let  $\langle j_1, \dots, j_k \rangle$ , with  $j_1 < \dots < j_k$ , symbolically denote a so-called *cylinder set* in  $X$  that is specified at sites  $j_i \in S$ , for  $1 \leq i \leq k$ . More precisely, these are sets of the form

$$(13) \quad \langle j_1, \dots, j_k \rangle = X_0 \times \dots \times X_{j_1-1} \times \{x_{j_1}\} \times [\dots] \times \{x_{j_k}\} \times X_{j_k+1} \times \dots \times X_n,$$

where  $[\dots]$  contains factors  $\{x_i\}$  or  $X_i$  depending on whether  $i$  appears in the set  $\{j_1, \dots, j_k\}$  or not. That is, with the cylinder set  $\langle j_1, \dots, j_k \rangle$ , we mean the set of types that have prescribed letters  $x_{j_1}, \dots, x_{j_k}$  at sites  $j_1, \dots, j_k$ , and arbitrary letters at all other sites. Note that the shorthand  $\langle j_1, \dots, j_k \rangle$  is symbolic in that it does not specify the letters explicitly (the  $x_{j_i}$  appear on the right-hand side of (13), but not on the left). Note also that the cylinder sets formalise our previous ‘ $*$ ’ notation:  $\omega_t(\langle j_1, \dots, j_k \rangle)$  is the marginal frequency of the types prescribed at sites  $j_1, \dots, j_k$ , and (1) may be rewritten as

$$\dot{\omega}_t(\langle S \rangle) = \sum_{\alpha \in L} \varrho_\alpha \left( \frac{1}{\|\omega_t\|} \omega_t(\langle 0, \dots, \lfloor \alpha \rfloor \rangle) \omega_t(\langle \lceil \alpha \rceil, \dots, n \rangle) - \omega_t(\langle S \rangle) \right).$$

Let us now define *linkage disequilibria* (or *correlation functions*) of order  $k$  via

$$(14) \quad d_t(\langle j_1, \dots, j_k \rangle) := (T_{\{\alpha < j_1\} \cup \{\alpha > j_k\}}(p_t))(\langle j_1, \dots, j_k \rangle).$$

That is, for a given cylinder set, LDEs emerge by applying, to  $p_t = \omega_t / \|\omega_t\|$ , the LDE operator defined by the links before the first and after the last element of that cylinder set, and evaluating the resulting quantity at the cylinder set. Clearly,  $d_t(\langle j_1, \dots, j_k \rangle)$  contains products of at most  $k$  marginal frequencies. Note that we have defined linkage disequilibria on the basis of the *normalised* quantities  $p_t$  for the sake of compatibility, and comparability, with related quantities in population genetics.

By (11) we know that, for every cylinder set  $\langle j_1, \dots, j_k \rangle$ ,

$$(15) \quad \frac{d}{dt} d_t(\langle j_1, \dots, j_k \rangle) = - \left( \sum_{j_1 < \alpha < j_k} \varrho_\alpha \right) d_t(\langle j_1, \dots, j_k \rangle);$$

we thus have linkage disequilibria *of all orders* that decouple and decay exponentially. It was shown in [6, p. 25] that the collection of  $d_t(\langle j_1, \dots, j_k \rangle)$ , for all index sets  $\{j_1, \dots, j_k\} \subset S$  and all choices of letters  $x_{j_i} \in X_{j_i}$ ,  $1 \leq i \leq k$ , is complete in that it uniquely determines  $p_t$ .

*An example.* Consider  $S = \{0, 1, 2, 3\}$ . The highest (fourth-order) LDE is, on the basis of (10) and (14), given by

$$\begin{aligned} d_t(\langle 0, 1, 2, 3 \rangle) &= (T_\emptyset(p_t))(\langle 0, 1, 2, 3 \rangle) = \sum_{H \subset L} (-1)^{|H|} (R_H(p_t))(\langle 0, 1, 2, 3 \rangle) \\ &= [0, 1, 2, 3] - [0][1, 2, 3] - [0, 1][2, 3] - [0, 1, 2][3] \\ &\quad + [0][1][2, 3] + [0][1, 2][3] + [0, 1][2][3] \\ &\quad - [0][1][2][3], \end{aligned}$$

where we have now used the shorthand  $[j_1, \dots, j_k] := p_t(\langle j_1, \dots, j_k \rangle)$ . By (15), we have  $\dot{d}_t(\langle 0, 1, 2, 3 \rangle) = -(\varrho_{1/2} + \varrho_{3/2} + \varrho_{5/2}) d_t(\langle 0, 1, 2, 3 \rangle)$ .

The third-, second-, and first-order LDEs are given by

$$\begin{aligned} d_t(\langle j_1, j_2, j_3 \rangle) &= [j_1, j_2, j_3] - [j_1][j_2, j_3] - [j_1, j_2][j_3] + [j_1][j_2][j_3], \\ d_t(\langle j_1, j_2 \rangle) &= [j_1, j_2] - [j_1][j_2], \\ d_t(\langle j_1 \rangle) &= [j_1], \end{aligned}$$

where the latter correspond to the letter frequencies, as usual.

Obviously, the  $d_t$  are correlation functions that measure the dependence between sites, at the various orders. It is important to note that they agree with other measures of LDE (see [9, Ch. V.4.2] for an overview) only up to order two. From order three onwards, they are special in that terms like  $[0, 2][1]$  or  $[0, 3][1, 2]$  are absent. This is due to the fact that our  $d_t$  are based on *ordered* partitions of  $S$ , and thus only contain terms ‘produced’ by composite recombinators; they are therefore adapted to single-crossover dynamics. In contrast, conventional LDEs or central moments are based on *unordered* partitions, which cannot be produced by composite recombinators.

### 3. THE STOCHASTIC MODEL AND ITS SIMULATION

The finite population counterpart of our deterministic model is the Moran model with single-crossover recombination. To simplify matters (and in order to clearly dissect the individual effects of recombination and resampling), we shall assume that resampling (traditionally referred to as genetic drift) and recombination occur independently of each other. More precisely, we assume a finite population of fixed size  $N$ , in which every individual experiences, independently of the others,

- resampling at rate  $b/2$ . The individual reproduces, the offspring inherits the parent’s type and replaces a randomly chosen individual (possibly its own parent).
- recombination at (overall) rate  $\varrho_\alpha$  at link  $\alpha \in L$ . Every individual picks a random partner (maybe itself) at rate  $\varrho_\alpha/2$ , and the pair exchanges the sites after link  $\alpha$ . That is, if the recombining individuals have types  $x$  and  $y$ , they are replaced by



the two offspring individuals  $(x_{<\alpha}, y_{>\alpha})$  and  $(y_{<\alpha}, x_{>\alpha})$ , as in the deterministic case, and Fig. 2. That is, the genetic material is conserved, and any resampling effect is excluded. As before, the per-capita rate of recombination at link  $\alpha$  is then  $\varrho_\alpha$ , because both orderings of the individuals lead to the same event. To avoid degeneracies, we shall assume throughout that  $\varrho_\alpha > 0$  for all  $\alpha \in L$ .

As in the deterministic model, the factor  $1/2$  arises every time *ordered* pairs (of an active individual and a randomly chosen partner) are considered (it is the same factor that also occurs when the Moran and Wright-Fisher models are compared [12, p. 23]). Note that the randomly chosen second individual (for resampling or recombination) may be the active individual itself; then, effectively, nothing happens. One might, for biological reasons, prefer to exclude these events by sampling from the remaining population only; but this means nothing but a change in time scale of order  $1/N$ .

To formalise this verbal description of the process, let the state of the population at time  $t$  be given by the collection  $Z_t = (Z_t(x))_{x \in X} \in E := \{z \in \{0, 1, \dots, N\}^{|X|} \mid \sum_x z(x) = N\}$ , where  $Z_t(x)$  is the number of individuals of type  $x$  at time  $t$ ; clearly,  $\sum_{x \in X} Z_t(x) = N$ . We also use  $Z_t$  in the sense of a (random counting) measure, in analogy with  $\omega_t$  (but keep in mind that  $Z_t$  is integer-valued and counts single individuals, whereas  $\omega_t$  denotes continuous frequencies in an infinite population). The letter  $z$  will be used to denote realisations of  $Z_t$  — but note that the symbols  $x, y$ , and  $z$  are not on equal footing ( $x$  and  $y$  will continue to be types). The stochastic process  $\{Z_t\}_{t \geq 0}$  is the continuous-time Markov chain on  $E$  defined as follows. If the current state is  $Z_t = z$ , two types of transitions may occur:

$$\begin{aligned}
 &\text{resampling:} \quad z \rightarrow z + s(x, y), \quad s(x, y) := \delta_x - \delta_y, \\
 (16) \quad &\quad \quad \quad \text{at rate } \frac{1}{2N} b z(x) z(y) \text{ for } (x, y) \in X \times X \\
 &\text{recombination:} \quad z \rightarrow z + r(x, y, \alpha), \\
 &\quad \quad \quad r(x, y, \alpha) := \delta_{(x_{<\alpha}, y_{>\alpha})} + \delta_{(y_{<\alpha}, x_{>\alpha})} - \delta_x - \delta_y, \\
 (17) \quad &\quad \quad \quad \text{at rate } \frac{1}{2N} \varrho_\alpha z(x) z(y) \text{ for } (x, y) \in X \times X, \alpha \in L
 \end{aligned}$$

(where  $\delta_x$  is the point measure on  $x$ , as before). Note that, in (16) and (17), transitions that leave  $E$  are automatically excluded by the fact that the corresponding rates vanish.

Note that ‘empty transitions’ ( $s(x, y) = 0$  or  $r(x, y, \alpha) = 0$ ) are explicitly included (they occur if  $x = y$  in resampling or recombination, and if  $x_{<\alpha} = y_{<\alpha}$  or  $x_{>\alpha} = y_{>\alpha}$  in recombination). This is convenient since the total reproduction and recombination rates in the population (as based on active individuals) are then given by  $bN/2$  and  $\varrho N/2$ , respectively, independently of  $z$ , a fact that comes in handy in the simulations. Since we only require  $Z_t$  at (equidistant) epochs  $t_i = i\Delta t$ ,  $i = 0, 1, 2, \dots$ , no waiting times for individual events need be generated; rather, we may draw the total number of events in a time interval  $\Delta t$  from the Poisson distribution with parameter  $N(b + \varrho)\Delta t/2$ . The nature of each event is then determined in the obvious way (the active individual is of type  $x$  with probability  $z(x)/N$ ; the second individual (chosen randomly for resampling or recombination) is of type  $y$  with probability  $z(y)/N$ ; the pair performs resampling with probability  $b/(b + \varrho)$ , or recombination with probability  $\varrho/(b + \varrho)$ ; if recombination occurs, then at link  $\alpha$  with probability  $\varrho_\alpha/\varrho$ ).

For simulations, the description of the process through (16) and (17) is perfectly adequate. For the theoretical analysis, however, one also needs the rates  $Q(z, z + v)$  for the transitions  $z \rightarrow z + v$  for *given*  $z \in E$ ,  $v \in E - z$  (where  $E - z := \{v \mid z + v \in E\}$ ). Here, one must take

care of the fact that distinct combinations of  $x, y$ , and  $\alpha$  may lead to the same ‘net’ transition. Clearly,  $Q(z, z + v) = Q^{(s)}(z, z + v) + Q^{(r)}(z, z + v)$ , where

$$(18) \quad Q^{(s)}(z, z + v) := \frac{b}{2N} \sum_{\substack{(x,y) \in X \times X: \\ s(x,y)=v}} z(x)z(y),$$

$$(19) \quad Q^{(r)}(z, z + v) := \frac{1}{2N} \sum_{\substack{(x,y) \in X \times X, \alpha \in L: \\ r(x,y,\alpha)=v}} \varrho_\alpha z(x)z(y).$$

Of course, the empty sum corresponds to impossible transitions and is understood as 0. Note that the sum in (18) contains more than one term only in the case  $v = 0$ ; however, (19) is a ‘true’ sum (of at least two terms) whenever  $v$  is of the form  $r(x, y, \alpha)$  for some  $x, y, \alpha$ .

Note that, with the rates for ‘empty transitions’ as ascribed to the diagonal elements  $Q(z, z)$ , the collection  $(Q(z, z'))_{z, z' \in E}$  does *not* form a Markov generator; but the corresponding generator  $\tilde{Q} = (\tilde{Q}(z, z'))_{z, z' \in E}$  is obtained by defining  $\tilde{Q}(z, z') = Q(z, z')$ , for  $z' \neq z$ , together with  $\tilde{Q}(z, z) = -\sum_{z': z' \neq z} Q(z, z')$ .

This rather informal definition of the Markov chain in terms of its transitions and their rates is all that is required for this article – after all, we are in the well-behaved situation of a finite state space. Excellent overviews of many aspects of Markov chains in continuous time (including the formal details) can be found in [20, Ch. 2] or [2, Ch. I.8].

In contrast to the situation in infinite populations, it now matters whether one or two offspring are created at a time. For comparison, we shall, therefore, also consider the alternate (biologically more realistic) recombination scheme in which an individual recombines with a random partner, but only *one* of the offspring is kept (chosen randomly), and replaces *one* of its parents (again chosen randomly). That is, instead of (17), we now have four recombination transitions

$$(20) \quad z \rightarrow z - \delta_x + \delta_{(x_{<\alpha}, y_{>\alpha})}, \quad z \rightarrow z - \delta_y + \delta_{(y_{<\alpha}, x_{>\alpha})},$$

$$(21) \quad z \rightarrow z - \delta_x + \delta_{(y_{<\alpha}, x_{>\alpha})}, \quad z \rightarrow z - \delta_y + \delta_{(x_{<\alpha}, y_{>\alpha})},$$

each at rate

$$(22) \quad \frac{1}{8N} \varrho_\alpha z(x)z(y), \text{ for } (x, y) \in X \times X, \alpha \in L.$$

Note that, in contrast to (17), the symmetry between  $x$  and  $y$  is broken in single transitions. Furthermore, only half as many replacements happen here; the corresponding deterministic dynamics is, therefore, given by (6) with  $\varrho_\alpha$  replaced by  $\varrho_\alpha/2$  for all  $\alpha \in L$ . More importantly, the recombination scheme (20), (21) no longer conserves the genetic material (since net replacements (e.g. of  $x_{>\alpha}$  by  $y_{>\alpha}$ ) take place), which introduces a (minor) resampling effect not present in (17). If not stated otherwise, however, we will stick to (17) for our recombination transitions.

#### 4. CONNECTIONS BETWEEN STOCHASTIC AND DETERMINISTIC MODELS

Let us now explore the connection between the stochastic process  $\{Z_t\}_{t \geq 0}$  on  $E$ , its normalised version  $\{\hat{Z}_t\}_{t \geq 0} = \{Z_t\}_{t \geq 0}/N$  on  $E/N$ , and the solution  $\omega_t = \varphi_t(\omega_0)$  of the differential equation (9). We shall first consider the case without resampling (i.e.,  $b = 0$ ) and show that  $\mathbb{E}(Z_t) = \varphi_t(Z_0)$  (and, consequently,  $\mathbb{E}(\hat{Z}_t) = \varphi_t(\hat{Z}_0)$ ), for *arbitrary*  $N$ . For  $b > 0$ , still with

finite  $N$ , this is no longer true; however,  $\widehat{Z}_t$  converges to  $\varphi_t(p_0)$  for  $N \rightarrow \infty$ , provided  $\widehat{Z}_0$  converges to  $p_0$ . We shall make this convergence explicit, and investigate it by means of simulations.

**4.1. General properties of the Moran model with recombination.** For lack of reference and the sake of completeness, we start by making explicit a plausible – and often implicitly-used – fact concerning the evolution of the expectation in finite-state continuous-time Markov chains.

**Fact 1.** *Let  $\{Z_t\}_{t \geq 0}$  be a continuous-time Markov chain on a finite subset  $E$  of  $\mathbb{Z}^d$ , as defined by its transitions from state  $z$  to state  $z + v$  ( $z, z + v \in E$ ) taking place at rates  $Q(z, z + v)$ . For all  $t \geq 0$ , the expectation of  $Z_t$  then satisfies the equation*

$$(23) \quad \frac{d}{dt} \mathbb{E}(Z_t) = \mathbb{E}(F(Z_t)),$$

where, for  $z \in E$ ,

$$(24) \quad F(z) := \sum_{v \in E - z} v Q(z, z + v).$$

Before turning to the proof, let us note that  $F(z)$  may be interpreted as the ‘mean rate-of-change vector’ of the chain in state  $z$ . A little later, we shall turn (23) into a differential equation for  $\mathbb{E}(Z_t)$ .

*Proof.* Let us first recall the elementary fact that, for the Markov chain starting in  $Z_0 = s$ , we have  $\mathbb{P}(Z_t = z) = P_t(s, z)$ , where we have suppressed dependence on  $s$  on the left-hand side, and  $P_t = (P_t(s, z))_{s, z \in E} = e^{t\tilde{Q}}$  is the Markov semigroup corresponding to  $\tilde{Q}$ , the generator of the chain (obtained from the collection of transition rates  $(Q(z, z + v))_{z, z + v \in E}$ , see above). Therefore,  $\mathbb{E}(Z_t) = \sum_{z \in E} z P_t(s, z)$ , and, since  $\frac{d}{dt} P(t) = P(t)\tilde{Q}$ ,

$$\begin{aligned} \frac{d}{dt} \mathbb{E}(Z_t) &= \sum_{z' \in E} z' \frac{d}{dt} P_t(s, z') = \sum_{z, z' \in E} z' P_t(s, z) \tilde{Q}(z, z') \\ &= \sum_{z, z' \in E} (z' - z) P_t(s, z) \tilde{Q}(z, z') = \sum_{z \in E} \sum_{v \in E - z} v P_t(s, z) \tilde{Q}(z, z + v) \\ &= \mathbb{E} \left( \sum_{v \in E - Z_t} v \tilde{Q}(Z_t, Z_t + v) \right) = \mathbb{E}(F(Z_t)), \end{aligned}$$

where the third equality sign is due to the fact that  $\tilde{Q}$  is a Markov generator.  $\square$

After this digression, let us return to the Moran model (with recombination *and* resampling), and show that  $F = \Phi$  on  $E$ , where  $\Phi$  continues to be the right-hand side of (6).

To this end, recall that  $Z_t$  is a (counting) measure, to which our projection operators may be applied in the usual way. In agreement with (3), we write  $\pi_I Z_t = Z_t \circ \pi_I^{-1}$ , i.e.,  $\pi_I Z_t$  is the marginal of  $Z_t$  w.r.t. the sites in  $I$ ; and, likewise, for a realisation  $z$  of  $Z_t$ . Again, we also use shorthands such as  $Z_t(x_{<\alpha}, *)$  to denote marginal frequencies (in this case,  $(\pi_{<\alpha} Z_t)(x_{<\alpha})$ ). Furthermore, we set  $E_I := \pi_I E$  for  $I \subset S$ .

To calculate  $F$ , let us write  $F = F^{(s)} + F^{(r)}$ , where  $F^{(s)}$  and  $F^{(r)}$  take care of transitions due to resampling and recombination, respectively. By symmetry of the transition rates for resampling,  $F^{(s)} = 0$  (type frequencies are martingales under resampling alone); we can therefore restrict ourselves to the recombination transitions defined by  $Q^{(r)}(z, z + v)$  of (19).

Summing, as in (19), over all possibilities for the gain or loss of a single  $x$ -individual (cf. Fig. 2, lower panel), one obtains for the  $x$ -component of  $F$

$$\begin{aligned}
 F_x(z) &= F_x^{(r)}(z) = \sum_{\substack{v \in E-z: \\ v(x)=\pm 1}} v(x) Q^{(r)}(z, z+v) \\
 (25) \quad &= \sum_{\alpha \in L} \frac{1}{N} \varrho_\alpha (z(x_{<\alpha}, *) z(*, x_{>\alpha}) - z(x_{<\alpha}, x_{>\alpha}) z(*, *)),
 \end{aligned}$$

where, of course,  $z(x_{<\alpha}, x_{>\alpha}) = z(x)$ , and  $z(*, *) = N$ . One immediately recognises the familiar structure of (1). We may therefore conclude:

**Fact 2.** *In the Moran model with resampling and recombination transitions according to (16) and (17), we have  $F = \Phi$  on  $E$ , with  $\Phi$  of (6).*

**4.2. Recombination without resampling: expectations for finite  $N$ .** Let us now return to Eq. (23). Note that, in general, it does not lead to a “closed” differential equation for  $\mathbb{E}(Z_t)$ , because it is not clear whether  $\mathbb{E}(F(Z_t))$  can be written as a function of  $\mathbb{E}(Z_t)$  alone. Clearly,  $\mathbb{E}(F(Z_t)) = F(\mathbb{E}(Z_t))$  if  $F$  is linear (or affine), as, for example, in Markov branching processes, or in the Moran model with mutation, but without recombination. But for nonlinear  $F$ , this tends to be violated, as nicely illustrated in [16, Ch. 1.4] for the case of the Ricker model in ecology. If  $F$  is nonlinear but polynomial (the usual case in population genetics or chemical reaction systems, for example), (23) can still serve as a starting point for an expansion involving a hierarchy of moments, which can be closed by suitable approximation methods (like, for example, the quasi-linkage equilibrium approach in [7]).

Our aim in this paragraph is to show that recombination *without* resampling is another (and apparently new) exception: despite its nonlinearity,  $\Phi$  satisfies  $\mathbb{E}(\Phi(Z_t)) = \Phi(\mathbb{E}(Z_t))$ , provided the marginals of  $Z_0$  are independent, which always applies if  $Z_0$  is fixed. This will require a lemma concerning the independence of marginal processes. For  $I \subset S$ , we introduce the ‘stretch’

$$J(I) := \{i \in S \mid \min(I) \leq i \leq \max(I)\},$$

and look at the projection of the recombination process on non-overlapping stretches. This is the content of the following lemma.

**Lemma 1.** *Let  $\{Z_t\}_{t \geq 0}$  be the recombination process without resampling as defined by the transition rates (17). Let  $A, B \subset S$  with  $J(A) \cap J(B) = \emptyset$ . Then,  $\{\pi_A \cdot Z_t\}_{t \geq 0}$  and  $\{\pi_B \cdot Z_t\}_{t \geq 0}$  are conditionally (on  $Z_0$ ) independent Markov chains on  $E_A$  and  $E_B$ .*

*Proof.* Clearly, for any given  $I \subset S$ ,  $\{\pi_I \cdot Z_t\}_{t \geq 0}$  is a stochastic process on  $E_I$ . Let us first show that  $\pi_A \cdot Z_t$  and  $\pi_B \cdot Z_t$  are individually Markov chains, and then establish that they are (conditionally) independent. For the first step, observe that, when  $Z_t$  performs the transition  $z \rightarrow z + v$  (on  $E$ ), with  $v = \delta_{(x_{<\alpha}, y_{>\alpha})} + \delta_{(y_{<\alpha}, x_{>\alpha})} - \delta_x - \delta_y$  for some  $\alpha \in L$  and  $x, y \in X$ , then  $\pi_I \cdot Z_t$  goes from  $\pi_I \cdot z \rightarrow \pi_I \cdot (z + v)$  on  $E_I$ , where  $\pi_I \cdot v = \delta_{\pi_I(x_{<\alpha}, y_{>\alpha})} + \delta_{\pi_I(y_{<\alpha}, x_{>\alpha})} - \delta_{\pi_I(x)} - \delta_{\pi_I(y)}$ . The rate for a corresponding ‘projected transition’  $v_I \in E_I - z_I$  ( $z_I \in E_I$ ) is then given by summing all rates of the original process that lead to the transition  $v_I$  under the projection. That is, with  $r_I(x_I, y_I, \alpha) := \delta_{(x_{I<\alpha}, y_{I>\alpha})} + \delta_{(y_{I<\alpha}, x_{I>\alpha})} - \delta_{x_I} - \delta_{y_I}$  and the

shorthand  $I_{<\alpha} := I \cap \{0, 1, \dots, \lfloor \alpha \rfloor\}$  (and likewise for  $I_{>\alpha}$ ), one gets

$$\sum_{v: \pi_I \cdot v = v_I} Q^{(r)}(z, z + v) = \frac{1}{2N} \sum_{\substack{x_I, y_I \in X_I, \alpha \in L: \\ r_I(x_I, y_I, \alpha) = v_I}} \varrho_\alpha((\pi_I \cdot z)(x_I))((\pi_I \cdot z)(y_I)),$$

for  $z \in E$ . Since these rates only depend on  $\pi_I \cdot z$  (rather than  $z$  itself), they define a collection of rates  $Q_I^{(r)}(z_I, z_I + v_I)$  ( $z_I \in E_I, v_I \in E_I - z_I$ ) so that  $\sum_{v: \pi_I \cdot v = v_I} Q^{(r)}(z, z + v) = Q_I^{(r)}(\pi_I \cdot z, \pi_I \cdot (z + v))$  for all  $z \in E$ . Therefore, the marginalised process  $\{\pi_I \cdot Z_t\}_{t \geq 0}$  is a Markov chain on  $E_I$  with transitions  $z_I \rightarrow z_I + v_I$  at rates  $Q_I^{(r)}(z_I, z_I + v_I)$ . (This is an example of the so-called *lumping procedure* for Markov chains, compare [10] and [17] for the general case, or [4] for the sequence context considered here). Our processes  $\{\pi_A \cdot Z_t\}_{t \geq 0}$  and  $\{\pi_B \cdot Z_t\}_{t \geq 0}$  are therefore Markov on  $E_A$  and  $E_B$ , respectively.

For the second step, we note that, for given  $\alpha$  and  $I$ , a net transition in  $\pi_I \cdot Z_t$  (i.e.,  $\pi_I \cdot v \neq 0$ ) requires  $\lfloor \alpha \rfloor, \lceil \alpha \rceil \in J(I)$ . Since  $J(A) \cap J(B) = \emptyset$  by assumption,  $\pi_A \cdot v \neq 0$  implies  $\pi_B \cdot v = 0$  and vice versa. A transition in  $\{\pi_A \cdot Z_t\}_{t \geq 0}$  therefore leaves  $\{\pi_B \cdot Z_t\}_{t \geq 0}$  unchanged, and vice versa. The joint process  $W_t := \{\pi_A \cdot Z_t, \pi_B \cdot Z_t\}_{t \geq 0}$  therefore has generator  $\tilde{Q}_{A,B} = \tilde{Q}_A \otimes \mathbb{1}_B + \mathbb{1}_A \otimes \tilde{Q}_B$ , where  $\tilde{Q}_A$  is the generator of  $\{\pi_A \cdot Z_t\}_{t \geq 0}$ ,  $\tilde{Q}_B$  the generator of  $\{\pi_B \cdot Z_t\}_{t \geq 0}$ , and  $\mathbb{1}_A$  and  $\mathbb{1}_B$  denote the identity on  $E_A$  and  $E_B$ , respectively. Hence, the marginal processes are conditionally independent, and the claim follows.  $\square$

**Remark 1.** Although we have, for ease of notation, formulated the above result for two subsets  $A$  and  $B$  only, the proof obviously goes through for any collection  $A_1, \dots, A_K \subset S$ , if the  $J(A_k)$ ,  $1 \leq k \leq K$ , are pairwise disjoint. We then obtain that  $\{\pi_{A_1} \cdot Z_t\}_{t \geq 0}, \dots, \{\pi_{A_K} \cdot Z_t\}_{t \geq 0}$  are conditionally independent Markov chains. If, furthermore,  $\pi_{A_1} \cdot Z_0, \dots, \pi_{A_K} \cdot Z_0$  are independent, conditional independence of the marginal processes turns into independence. In particular, this is the case if  $Z_0$  is fixed (since  $\mathbb{P}(Z_0 \in \cdot)$  is then a point measure on some  $s \in E$ ).

As an immediate consequence, we now arrive at

**Theorem 1.** *Let  $\{Z_t\}_{t \geq 0}$  be the recombination process without resampling (i.e.,  $b = 0$ ), and let  $Z_0$  be fixed. Then,  $\mathbb{E}(Z_t)$  satisfies the differential equation*

$$\frac{d}{dt} \mathbb{E}(Z_t) = \Phi(\mathbb{E}(Z_t))$$

with initial value  $Z_0$ , and  $\Phi$  from (6); therefore,

$$\mathbb{E}(Z_t) = \varphi_t(Z_0), \quad \text{for all } t \geq 0,$$

with  $\varphi_t$  from (9).

*Proof.* Applying Fact 1, Eq. (4), Lemma 1, and the linearity of the expectation, one finds

$$\begin{aligned}
\frac{d}{dt}\mathbb{E}(Z_t) &= \sum_{\alpha \in L} \varrho_\alpha \mathbb{E}((R_\alpha - \mathbb{1})Z_t) \\
&= \sum_{\alpha \in L} \frac{\varrho_\alpha}{N} [\mathbb{E}((\pi_{<\alpha}.Z_t) \otimes (\pi_{>\alpha}.Z_t)) - \mathbb{E}(Z_t)] \\
&= \sum_{\alpha \in L} \frac{\varrho_\alpha}{N} [\mathbb{E}(\pi_{<\alpha}.Z_t) \otimes \mathbb{E}(\pi_{>\alpha}.Z_t) - \mathbb{E}(Z_t)] \\
&= \sum_{\alpha \in L} \frac{\varrho_\alpha}{N} [(\pi_{<\alpha}.\mathbb{E}(Z_t)) \otimes (\pi_{>\alpha}.\mathbb{E}(Z_t)) - \mathbb{E}(Z_t)] \\
&= \Phi(\mathbb{E}(Z_t)).
\end{aligned}$$

(Clearly, the third step is the decisive one; it hinges on the independence of the marginals.)  $\square$

**Remark 2.** Note that Theorem 1 does *not* hold for the alternate recombination scheme (22). This is because the associated resampling effect already destroys the validity of Lemma 1, which is essential for the proof. We will come back to this in the next paragraph; but let us already mention here that the deterministic solution continues to be an excellent approximation to the expectation, as long as recombination rates are small. It will also be shown below that, in practice, the resampling effect thus introduced is minor as long as recombination rates are small, and averages continue to be very close to the deterministic solution.

Now that the equivalence between the deterministic solution and the expectation has been safely established at the level of the type frequencies, it immediately carries over to the linkage disequilibria:

**Corollary 1.** *Under the assumptions of Thm. 1, one has, for all  $t \geq 0$ ,*

$$\mathbb{E}(T_G Z_t) = T_G(\varphi_t(Z_0)).$$

*Proof.* Use the definitions of  $T_G$  and  $R_H$ , then Remark 1, and finally Theorem 1.  $\square$

In particular, if we define

$$(26) \quad C_t(\langle j_1, \dots, j_k \rangle) := \left( T_{\{\alpha < j_1\} \cup \{\alpha > j_k\}}(\widehat{Z}_t) \right) (\langle j_1, \dots, j_k \rangle)$$

(so that  $C_t$  is the stochastic equivalent of  $d_t$  of (14)), we obtain the relation  $\mathbb{E}(C_t(\langle j_1, \dots, j_k \rangle)) = d_t(\langle j_1, \dots, j_k \rangle)$ , for all cylinder sets  $\langle j_1, \dots, j_k \rangle$ .

Fig. 3 illustrates the findings of this paragraph for a four-locus two-allele system with recombination and no resampling. We have chosen  $\rho = 0.008$  and  $\rho_\alpha = \rho/3$  for  $\alpha \in \{\frac{1}{2}, \frac{3}{2}, \frac{5}{2}\} = L$ , which corresponds to four sites spaced evenly across a stretch of  $8 \cdot 10^5$  bp, at a per-nucleotide recombination rate of  $10^{-8}$ . The Figure shows haplotype frequencies and highest-order LDEs, both as single trajectories and averages over many realisations, as compared to the solution of the deterministic recombination equation. For very small populations ( $N = 100$ ), single trajectories fluctuate markedly, but averages over 100 realisations are already indistinguishable from the deterministic solution—in line with the property of the expectation just established. For larger populations, the stochasticity is already greatly reduced in single trajectories—but this is the topic of the next paragraph.

The Figure also contains results from the alternate resampling scheme with one offspring only, (20) – (22). Although Theorem 1 is not exactly valid in this case, the resampling effect thus introduced is obviously minor as long as recombination rates are small, and averages continue to be very close to the deterministic solution.

**4.3. Recombination and resampling: the infinite population limit.** Let us now include resampling, at rate  $b/2 > 0$ , and consider the stochastic process  $\{Z_t^{(N)}\}_{t \geq 0}$  defined by both (16) and (17), where we add the upper index here to indicate the dependence on  $N$ . Now,  $\Phi$  and  $\mathbb{E}$  no longer commute. The processes  $\{\pi_{<\alpha} \cdot Z_t\}_{t \geq 0}$  and  $\{\pi_{>\alpha} \cdot Z_t\}_{t \geq 0}$  are still individually Markov, but their resampling events are coupled (replacement of  $y_{<\alpha}$  by  $x_{<\alpha}$  is always tied to replacement of  $y_{>\alpha}$  by  $x_{>\alpha}$ ). Hence the marginal processes fail to be independent, so that no equivalent of Lemma 1 holds.

Let us, therefore, change focus and consider the normalised version  $\{\widehat{Z}_t^{(N)}\}_{t \geq 0} = \{Z_t^{(N)}\}_{t \geq 0}/N$ . It seems to be general folklore in population genetics that, in the limit  $N \rightarrow \infty$ , the relative frequencies of the Moran or Wright-Fisher model cease to fluctuate and are then given by the solution of the corresponding deterministic equation. This is implied no matter which evolutionary forces (like mutation, selection, or recombination) are included into the model; in our case,  $\{\widehat{Z}_t^{(N)}\}_{t \geq 0}$  should be described by the differential equation (1) as  $N \rightarrow \infty$ . However, the limit theorem behind this is usually not made explicit, and, in fact, does not seem to be well known in the population genetics literature. Indeed, it is given by a very general law of large numbers due to Ethier and Kurtz [13, Thm. 11.2.1], which provides the formal justification for a very large class of models in biology that are stochastic at the microscopic scale but are adequately described deterministically if the population size is large; they range from biochemical reaction kinetics to population dynamics and population genetics. For our special case, it reads as follows.

**Proposition 1.** *Consider the family of processes  $\{\widehat{Z}_t^{(N)}\}_{t \geq 0} = \frac{1}{N}\{Z_t^{(N)}\}_{t \geq 0}$ ,  $N = 1, 2, \dots$ , where  $\{Z_t^{(N)}\}_{t \geq 0}$  is defined by (16) and (17). Assume that the initial states are chosen so that  $\lim_{N \rightarrow \infty} \widehat{Z}_0^{(N)} = p_0$ . Then, for every given  $t \geq 0$ , one has*

$$(27) \quad \lim_{N \rightarrow \infty} \sup_{s \leq t} |\widehat{Z}_s^{(N)} - p_s| = 0$$

with probability 1, where  $p_s := \varphi_s(p_0)$  is the solution of the deterministic recombination equation (1).

*Proof.* To apply Thm. 11.2.1 of [13], we need a linear scaling of the transition rates with  $N$ . More precisely, we must show that the transition rates  $Q^{(N)}(z, z+v)$  of the process  $\{Z_t^{(N)}\}_{t \geq 0}$  are, for all  $z, z+v \in E$ , of the form

$$(28) \quad Q^{(N)}(z, z+v) = Nq_v(z/N), \quad \text{for all } N,$$

with non-negative functions  $q_v$  defined on a subset of  $\mathbb{R}_{\geq 0}^{|X|}$ . Setting

$$(29) \quad q_v^{(s)}(p) := \frac{b}{2} \sum_{\substack{(x,y) \in X \times X: \\ s(x,y)=v}} p(x)p(y),$$

$$(30) \quad q_v^{(r)}(p) := \frac{1}{2} \sum_{\substack{(x,y) \in X \times X, \alpha \in L: \\ r(x,y,\alpha)=v}} \varrho_\alpha p(x)p(y),$$

together with  $q_v := q_v^{(s)} + q_v^{(r)}$ , and recalling that

$$Q^{(N)}(z, z + v) = Q^{(N,s)}(z, z + v) + Q^{(N,r)}(z, z + v)$$

(from (18) and (19) – now with notational dependence on  $N$ ), it is obvious that (28) is indeed satisfied. (Observe that this just reflects the relation  $z(x)z(y)/N = N \frac{z(x)}{N} \frac{z(y)}{N}$ .) The normalised process  $\{\hat{Z}^{(N)}\}_{t \geq 0}$  on  $E/N$  has the corresponding transitions  $z/N \rightarrow z/N + v/N$ , again at rates  $Nq_v(z/N)$ . Thus, the collection of processes  $\{\hat{Z}^{(N)}\}_{t \geq 0}$  constitutes what is known as a *density-dependent family corresponding to the  $q_v$*  [13, p. 455].

Now, for such a density-dependent family, Thm. 11.2.1 of [13] implies (27) if  $p_t$  solves  $\dot{p}_t = f(p_t)$ , with initial value  $p_0$ , and  $f(p) := \sum_v v q_v(p)$ . Proceeding as in (24) and (25), we obtain  $f(p) = \Phi(p)$  in analogy with Fact 2.  $\square$

Note that the convergence in (27) applies for any given  $t$ , but need not hold for  $t \rightarrow \infty$ ; we shall come back to this point in the Discussion. Note also that the convergence carries over to linkage disequilibria, since  $T_G(\hat{Z}_t)$  converges to  $T_G(p_t)$  (in the above sense).

The question that remains is whether this limit result bears any relevance to real populations, which are always finite. How large must  $N$  be for the infinite-population limit to yield a reasonable approximation?

We have investigated this by means of simulations of our four-locus two-allele system (Fig. 4). As is to be expected, single realisations, as well as averages, approach the deterministic limit with increasing  $N$ ; and for  $N = 10^5$ , the stochastic process is already very close to the limit. This is observed at the level of type frequencies, as well as linkage disequilibria. The situation is very similar to that without resampling (Fig. 3, (a) and (e)), except that somewhat larger population sizes are required due to the increased stochasticity induced by resampling. In contrast, averages over *small* populations are not expected to, and in fact do not, get close to the deterministic solution.

## 5. DISCUSSION

The main purpose of this article has been to clarify some relationships between the Moran model with recombination and the deterministic recombination model (both with single crossovers). To separate the effects of recombination from those of resampling, we formulated the Moran model in its ‘decoupled’ version, with independent recombination and resampling events; this approach is also taken in [21], for example. The coupling of recombination to resampling (which happens to be biological reality) is thus neglected. Put differently, our model describes correctly the resampling effects due to reproduction events that do not involve recombination, but neglects those resampling events that occur in the course of reproduction associated with recombination. But recombination is rare (relative to reproduction), at least at the molecular level aimed at by the single-crossover model; and the bias introduced by this simplification is accordingly small, as also illustrated by our simulations of the alternative recombination scheme (20) – (22).

As the main result of this article, we have shown:

- (1) For recombination without resampling, the expected type frequencies are given by the deterministic dynamics, for arbitrary (even small) population sizes. Although this is not a biologically realistic situation, it yields insight into the Moran model with recombination, and establishes a relationship between finite and infinite populations that is somewhat unexpected in view of the inherent nonlinearity of recombination. The key



observation that led to this result rests upon the fact that the marginal processes (i.e., the type frequencies at the sites before resp. after a given link) are independent Markov chains. Of course, this is related to the fact that, in the absence of resampling, the genetic material is completely conserved (just rearranged); in particular, types cannot go to fixation.

- (2) The combined model *with* resampling deviates (in expectation) from the deterministic dynamics; in view of the findings above, these deviations are solely due to resampling, rather than recombination. But the infinite population limit continues to apply, and is, again, given by the deterministic recombination equation. We have investigated its range of validity here for one representative scenario (four biallelic loci, a time horizon of  $T = 300$ , an expected number of  $T\varrho_\alpha = 300 \cdot 0.008/3 = 0.8$  recombination events per link and individual, and  $Tb/2 = 300$  resampling events per individual); then, a population of moderate size ( $N = 10^5$ ) is already close to the deterministic limit. It should be noted, however, that this result is expected to further vary with:

- (a) The number of sites, or of alleles per site (or, more generally, the size of the type space): The deterministic limit can only be a good approximation if *all types* are well-populated; for a larger type space, the required population size will be larger. But situations with four loci (or variable nucleotide sites) already cover many interesting biological situations; and typing them (and determining the corresponding four-way LDEs) is, after all, a veritable task that yields a great deal of information. It remains to be investigated, however, how the quality of the approximation changes when there are more than two alleles per site.
- (b) The time horizon: The law of large numbers (27) holds for every given, finite time horizon, but need not carry over to  $t \rightarrow \infty$ . Indeed, if resampling is present, the population size required to get close to the deterministic solution is expected to grow over all bounds with increasing  $t$ . This is because, for every finite  $N$ , the Moran model with resampling and recombination is an *absorbing* Markov chain, which leads to fixation (i.e., a homogeneous population of uniform type) in finite time with probability one (for the special case of just two types without recombination, the expected time is of order  $N$ , if the initial frequencies are both  $1/2$  [14, p. 93]). In sharp contrast, the deterministic system never loses any type, and the stationary state, the complete product measure (12), is, in a sense, even the most variable state accessible to the system. For increasing  $N$ , finite populations stay close to the deterministic limit for an increasing length of time (see Fig. 4). Indeed, our main interest here is this time horizon during which substantial changes in LDE occur, and this is described by the deterministic model; in contrast, the deterministic solution does not provide a meaningful picture for the equilibrium situation. The case would be very different if mutation were included into the model, since this would turn the absorbing Markov chain into an ergodic one, whose stationary distribution allows a meaningful comparison with the deterministic dynamics even for  $t \rightarrow \infty$ .

Let us finally discuss implications for the corresponding model in discrete time, that is, the Wright-Fisher model with recombination. Again, we may consider

- a model without resampling (i.e., the only events are single crossovers between pairs of individuals, with pairs of offspring replacing pairs of parents): Although one might assume this model to behave like the corresponding discrete-time dynamical system for

finite  $N$ , *in expectation*, we indeed stipulate that this is not the case. The reason is that, in discrete time, the marginal processes (related to non-overlapping stretches) cease to be independent. This is because such independence would imply the possibility of two or more crossovers in one generation – in contrast to the single-crossover assumption. For the same reason, the deterministic single-crossover model in discrete time is much more difficult to solve than in continuous time, see the discussion in [6].

- recombination combined with resampling: It is strongly expected that, for  $N \rightarrow \infty$ , there is again a law of large numbers, which makes the relative frequencies converge to the corresponding deterministic dynamics, in analogy with Prop. 1. However, no explicit equivalent of the underlying Theorem 11.2.1 of [13] is known to the authors.

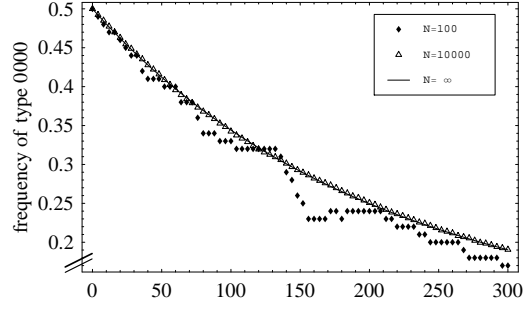
**Acknowledgments.** It is our pleasure to thank Richard Hudson for a decisive hint on the structure of stochastic recombination, and the ICMS workshop on Mathematical Genetics (Edinburgh 2006) for providing ample opportunity to discuss the results. We are grateful to Michael Baake, Matthias Birkner, and Thiemo Hustedt for critically reading the manuscript. This work was supported by the Dutch-German Bilateral Research Group on Random Spatial Models in Physics and Biology (DFG-FOR 498).

#### REFERENCES

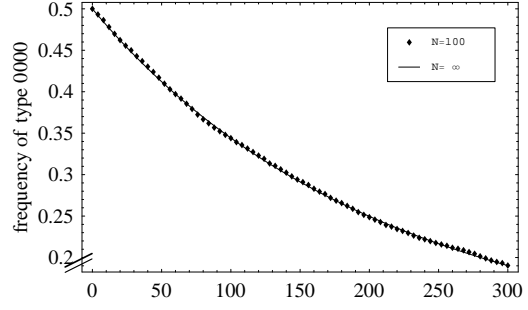
- [1] M. Aigner. *Combinatorial Theory*. Springer, Berlin, 1979. Reprint 1997.
- [2] S. Asmussen. *Applied Probability and Queues*. Springer, New York, 2nd edition, 2003.
- [3] E. Baake. Mutation and recombination with tight linkage. *J. Math. Biol.*, 42:455–488, 2001.
- [4] E. Baake, M. Baake, A. Bovier, and M. Klein. An asymptotic maximum principle for essentially nonlinear evolution models. *J. Math. Biol.*, 50:83–114, 2005. q-bio.PE/0311020.
- [5] M. Baake. Recombination semigroups on measure spaces. *Monatsh. Math.*, 146:267–278, 2005. math.CA/0506099.
- [6] M. Baake and E. Baake. An exactly solved model for mutation, recombination and selection. *Canad. J. Math.*, 55:3–41, 2003. math.CA/0210422.
- [7] N. Barton and M. Turelli. Natural and sexual selection on many loci. *Genetics*, 127:229–255, 1991.
- [8] J. Bennett. On the theory of random mating. *Ann. Hum. Genet.*, 18:311–317, 1954.
- [9] R. Bürger. *The Mathematical Theory of Selection, Recombination, and Mutation*. Wiley, Chichester, 2000.
- [10] C. Burke and M. Rosenblatt. A Markovian function of a Markov chain. *Ann. Math. Statist.*, 29:1112–1122, 1958.
- [11] K. Dawson. The evolution of a population under recombination: How to linearize the dynamics. *Linear Algebra Appl.*, 348:115–137, 2002.
- [12] R. Durrett. *Probability Models for DNA Sequence Evolution*. Springer, New York 2002.
- [13] S. N. Ethier and T. G. Kurtz. *Markov Processes – Characterization and Convergence*. Wiley, New York, 1986. Reprint 2005.
- [14] W. Ewens. *Mathematical Population Genetics*. Springer, Berlin, 2nd ed., 2004.
- [15] H. Geiringer. On the probability theory of linkage in Mendelian heredity. *Ann. Math. Statist.*, 15:25–57, 1944.
- [16] P. Haccou, P. Jagers and V.A. Vatutin. *Branching Processes – Variation, Growth, and Extinction of Populations*. Cambridge University Press, Cambridge, 2005.
- [17] J. G. Kemeny and J. L. Snell. *Finite Markov Chains*. Springer, New York, 1981.
- [18] Y. Lyubich. *Mathematical Structures in Population Genetics*. Springer, New York, 1992.
- [19] D. McHale and G. Ringwood. Haldane linearization of baric algebras. *J. London Math. Soc.*, 28:17–26, 1983.
- [20] J. R. Norris. *Markov Chains*. Cambridge University Press, Cambridge, 1997. reprint 1999.
- [21] P. Pfaffelhuber, B. Haubold, and A. Wakolbinger. Approximate genealogies under genetic hitchhiking. *Genetics*, 174:1995–2008, 2006.
- [22] E. Popa. Some remarks on a nonlinear semigroup acting on positive measures. In O. Carja and I. I. Vrabie, eds., *Applied Analysis and Differential Equations*, pp. 308–319. World Scientific, Singapore, 2007.

- [23] G. Ringwood. Hypergeometric algebras and Mendelian genetics. *Nieuw Archief v. Wiskunde*, 3:69–83, 1985.

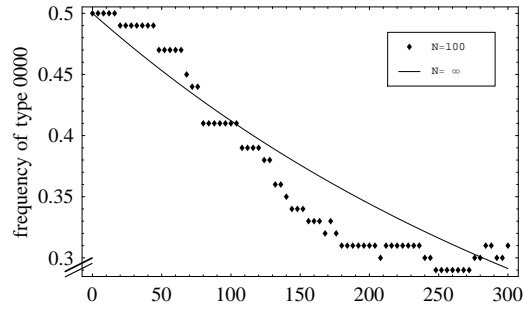
FACULTY OF TECHNOLOGY, BIELEFELD UNIVERSITY, 33594 BIELEFELD, GERMANY  
*E-mail address*: {ebaake, ihildebr}@techfak.uni-bielefeld.de



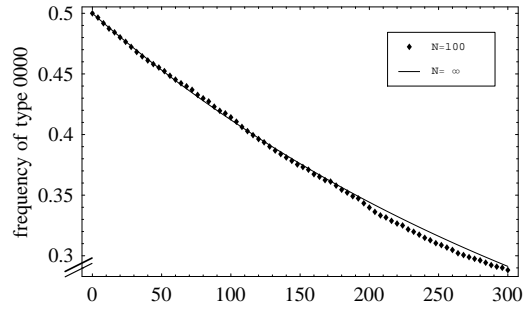
(a) haplotype frequency, single trajectories



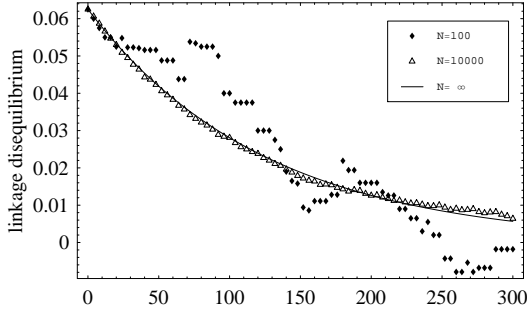
(b) haplotype frequency, averaged over 100 runs



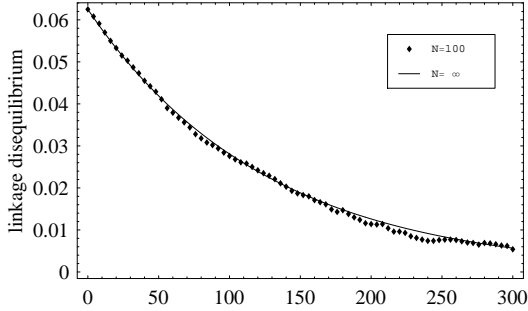
(c) as (a), but with one offspring only



(d) as (b), but with one offspring only



(e) LDE, single trajectories



(f) LDE, averaged over 100 runs

FIGURE 3. Recombination without resampling in a 4-locus 2-allele system. Sites:  $S = \{0, 1, 2, 3\}$ , types:  $X = \{0, 1\}^4$ ; recombination rates:  $\rho = 0.008$ ,  $\varrho_{1/2} = \varrho_{3/2} = \varrho_{5/2} = \rho/3$ ; initial conditions:  $Z_0(0000) = Z_0(1111) = N/2$  (no other types are present). Horizontal axis: time ( $t$ ). Panels (a) – (d) show realisations of  $\hat{Z}_t(0000)$  for various population sizes, as single trajectories (a), or means over 100 (b) realisations; also shown is  $p_t(0000) = (\varphi_t(\hat{Z}_0))(0000)$  (the deterministic solution or  $N \rightarrow \infty$  limit), which equals  $\mathbb{E}(\hat{Z}_t)$  by Thm. 1 (and is partly hidden between the diamonds in (b) and (d)). Clearly, the sample mean is well described by the expectation, where averaging is faster in larger populations. Panel (c) shows a single trajectory, and panel (d) an average over 100 realisations, for a simulation according to Eqs. (20)–(22), in which only one offspring individual is produced, rather than two as in Eq. (17), the assumption underlying all other simulations. Obviously, this only introduces a minor resampling effect and corresponding systematic deviation from the deterministic limit (but note that, relative to the usual resampling scheme, the dynamics is slowed down by a factor of  $1/2$ ). Panels (e) and (f) show  $C_t(\langle 0, 1, 2, 3 \rangle)$  (of Eq. (26)), for type 0000, and various population sizes, as single trajectories (e), or means (f) over 100 realisations, and compares them with the deterministic quantity (or  $N \rightarrow \infty$  limit)  $d_t(\langle 0, 1, 2, 3 \rangle)$ .

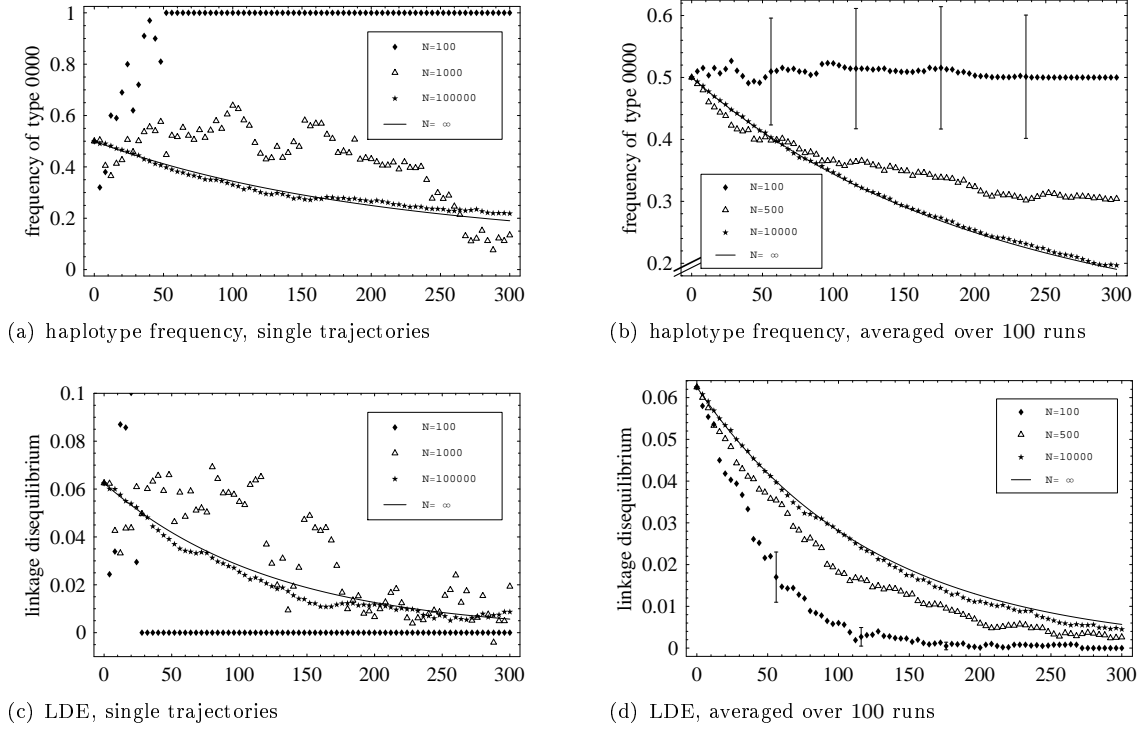


FIGURE 4. Recombination with resampling in the 4-locus 2-allele system. Sites, types, recombination rates, and initial conditions as in Fig. 3. Resampling rate:  $b = 2$ . Horizontal axis: time ( $t$ ). Panels (a) and (b) show realisations of  $\hat{Z}_t(0000)$  for various population sizes, as individual trajectories (a) or averages over 100 realisations (b); also shown is  $p_t(0000) = (\varphi_t(\hat{Z}_0))(0000)$  (the corresponding deterministic solution, or  $N \rightarrow \infty$  limit), partly hidden between the  $N = 10000$  stars in (b). Single trajectories (a) approach the infinite population limit already for moderate population sizes ( $N = 10^5$ ). In contrast, averages (b) deviate strongly from the deterministic solution if the population is small (error bars on the  $N = 100$  curve correspond to 95 % confidence intervals), but are hard to distinguish from the infinite population limit for  $N = 10000$ . Analogous results hold true for single trajectories (c), and averages over 100 realisations (d), of the highest-order LDE,  $C_t(\langle 0, 1, 2, 3 \rangle)$ , evaluated for type 0000; the  $N = \infty$  line is  $d_t$ .